Klassifikation in hochdimensionalen Räumen - Supportvektormaschinen

Vorlesung 8, Maschinelles Lernen

Dozenten: Prof. Dr. M. O. Franz, Prof. Dr. O. Dürr

HTWG Konstanz, Fakultät für Informatik

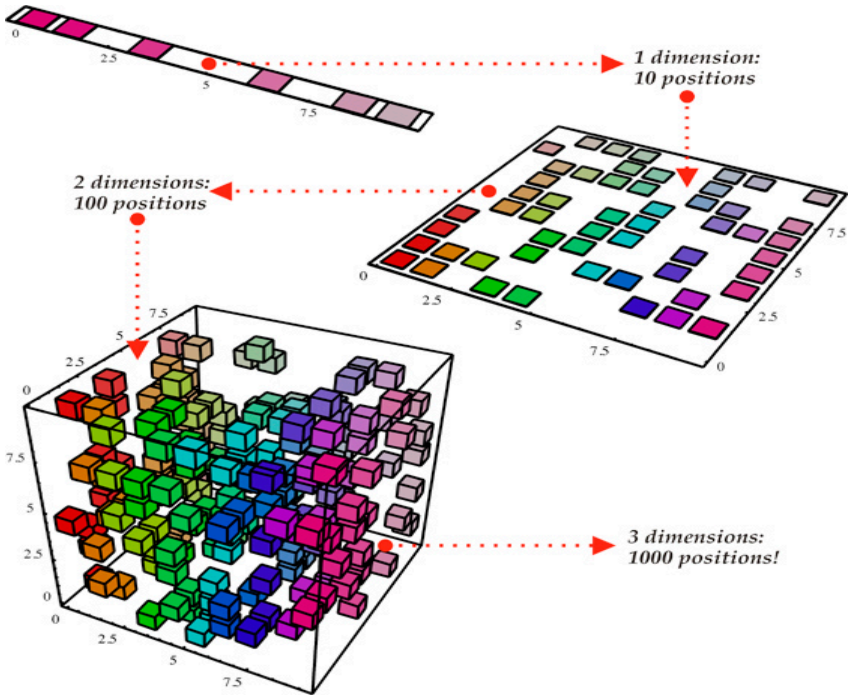
Übersicht

Übersicht

Fluch der Dimensionalität

Fluch der Dimensionalität: Die Komplexität von Funktionen mit mehreren Variablen kann exponentiell mit der Dimension wachsen, d.h. die Schätzung einer solchen Funktion erfordert ebenfalls exponentiell viele Daten.

Beispiel Dichteschätzung:



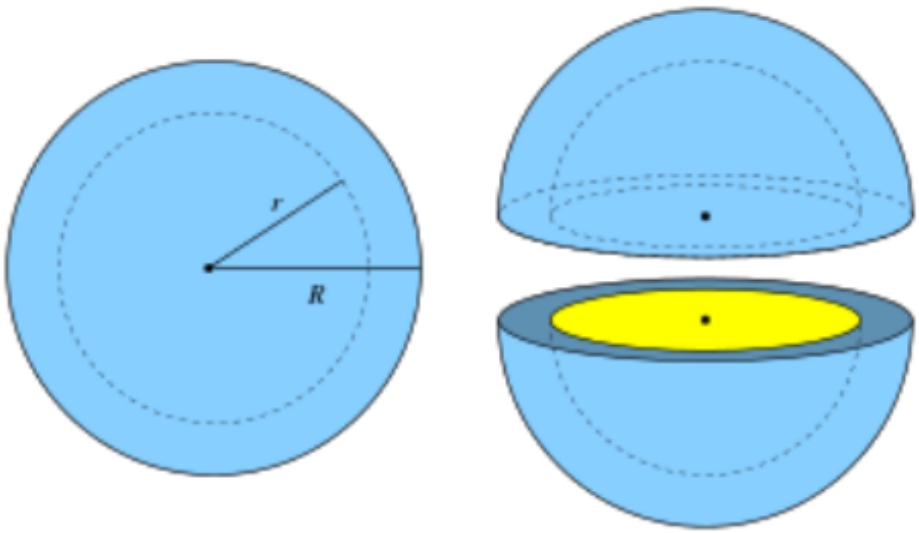
dim	n for 100% coverage
1	10
2	100
3	1000
6	10^6
p	10^p

[Tan, Steinbach, Kumar]

Die Anzahl der zum Abtasten einer hochdimensionalen Trennfläche benötigten Datenpunkte skaliert ebenso!

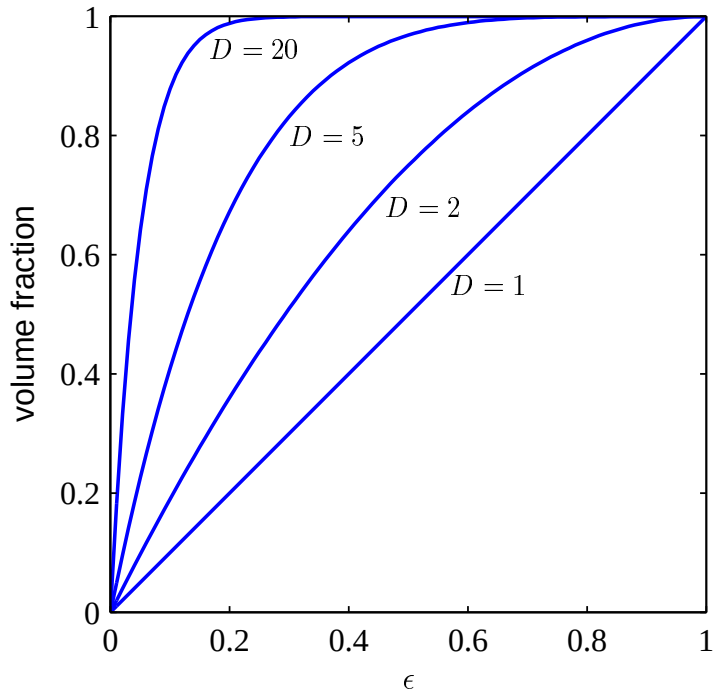
Hochdimensionale Daten sind hauptsächlich an der Oberfläche (1)

Wir nehmen an, dass die Daten annähernd gleichverteilt sind. Damit entsprechen die Anzahl der Beobachtungen dem Volumen.



Welcher Volumenanteil einer D-dimensionalen Kugel mit Radius R liegt in der Schale von r und R?

Hochdimensionale Daten sind hauptsächlich an der Oberfläche (2)



[Bishop, 2006]

Wir nehmen $R = 1$ und $r = 1 - \epsilon$ an.
Volumen der Kugel:

$$V(r) = k \cdot r^D$$

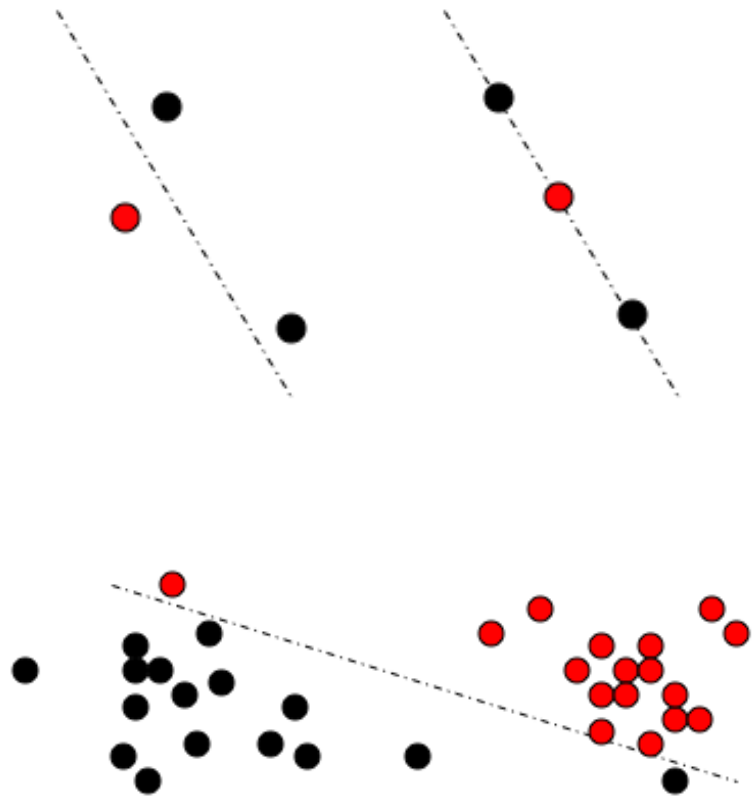
Anteil der Schale:

$$\frac{V(1) - V(1-\epsilon)}{V(1)} = 1 - (1-\epsilon)^d$$

$\Rightarrow$ In hochdimensionalen Räumen ist das Volumen einer Kugel in einer dünnen Schale an der Oberfläche konzentriert. Schält man eine hochdimensionale Orange, bleibt fast nichts übrig!

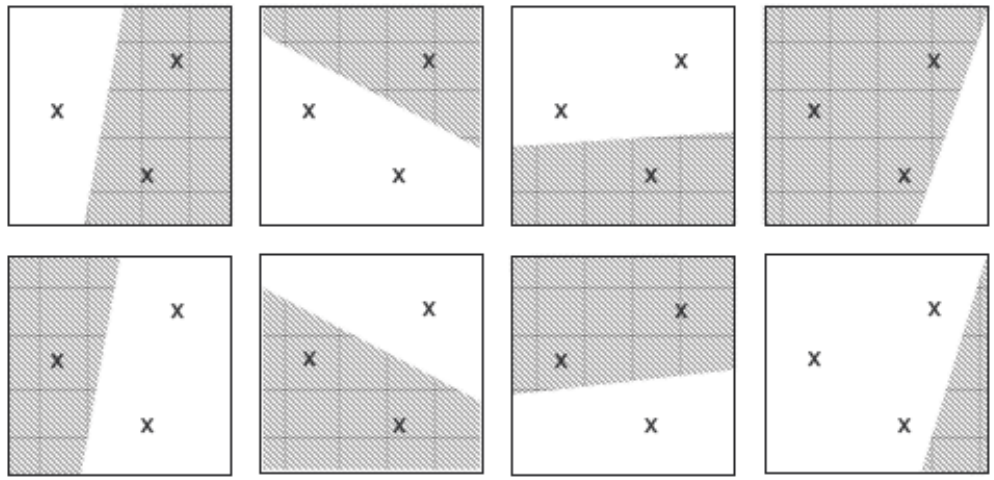
Lineare Trennbarkeit in höheren Dimensionen

- Bei d-dimensionalen Daten sind weniger als $d + 2$ Datenpunkte fast immer linear trennbar.
- Selbst wenn die Trainingsdaten linear trennbar sind, folgt daraus nicht, dass der resultierende Klassifikator gut auf unabhängigen Testdaten funktioniert.
- Man sollte also ein Mehrfaches von d an Trainingsdaten haben, um den Klassifikator zu überbestimmen und sicherzustellen, daß Trainings- und Testdaten ähnlich genug sind.



Wie kann man in einem hoch- oder unendlichdimensionalen Merkmalsraum klassifizieren, ohne zu viele Daten zu benötigen?

Klassifikation mit beliebigen Trennebenen



[B. Schölkopf]

- In hoch- und unendlichdimensionalen Räumen kann i.A. jede beliebige Punktekonfiguration getrennt werden.
- Dass genau die gleiche Aufteilung auch bei unabhängigen Testdaten gilt, ist sehr unwahrscheinlich ($\Rightarrow$ overfitting).

Klassifikation in hochdimensionalen Räumen

- Die Trennflächen zwischen den zu zwei Klassen gehörenden Punktewolken können i. A. immer komplexer mit zunehmender Dimension werden.
- Erforderlich wäre dazu eine exponentiell mit der Dimensionalität anwachsende Datenmenge $\Rightarrow$ i.A. ist damit Klassifikation in hoch-und unendlichdimensionalen Räumen unmöglich.
- Um dennoch lernen zu können, muss die Verteilung **gutartig** sein. Ein funktionierender Lernalgorithmus muss solche gutartigen Strukturen in den Daten erkennen und daraus Vorteile ziehen.
- Reale Probleme besitzen in fast allen Fällen eine gutartige Struktur.
- Perzeptronen beenden das Training, sobald die Trainingsdaten getrennt sind, d.h. hier gibt es keinen Mechanismus, der gutartige Strukturen berücksichtigt.

Gutartige Klassifikationsprobleme

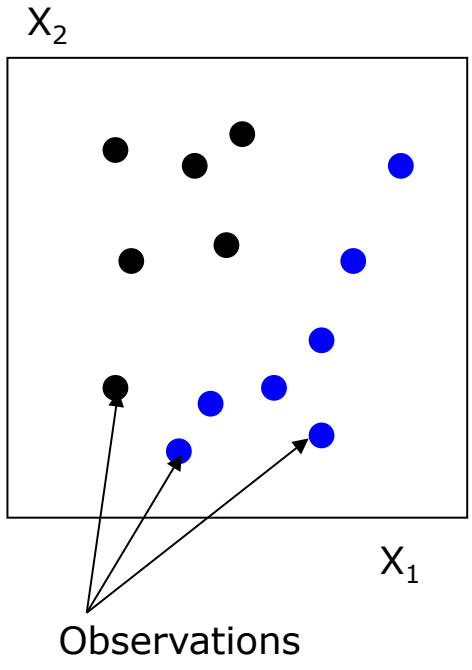
Die statistische Lerntheorie befaßt sich mit der Abschätzung des Generalisierungsfehlers von Lernmaschinen aus bestimmten Eigenschaften der Daten.

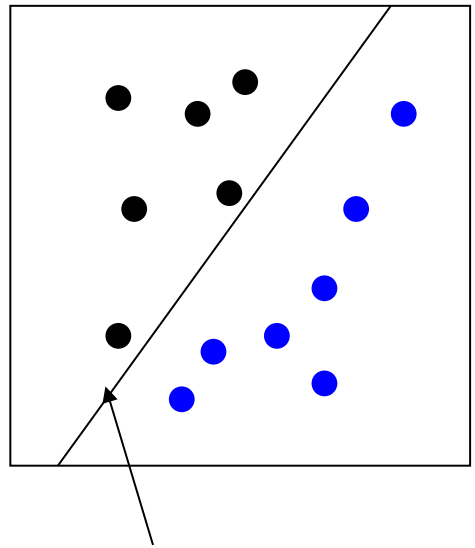
- Bei linearen Klassifikationsproblemen stellt sich heraus, daß linear trennbare Probleme mit hinreichend großer Trennbreite gutartig sind.
- Man kann beweisen, daß der Generalisierungsfehler mit hoher Wahrscheinlichkeit um so kleiner wird, je größer die Trennbreite ist. Erstaunlicherweise gilt dieses Resultat unabhängig von der Dimensionalität des Inputraumes.

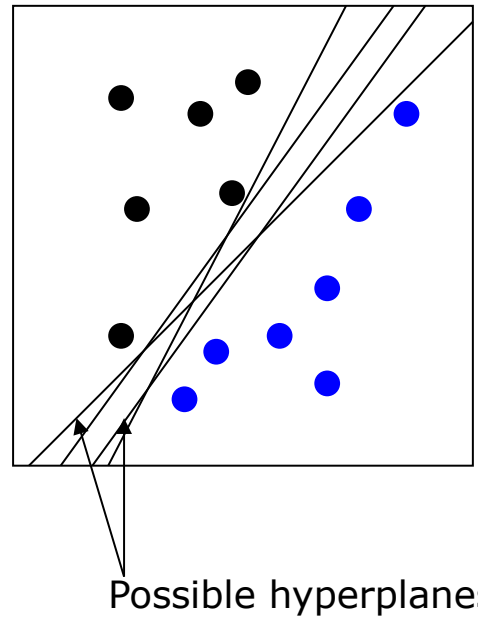
$\Rightarrow$ Sucht man statt einer beliebigen Trennebene solche mit einer möglichst großen Trennbreite, dann kann man in Räumen beliebiger Dimension klassifizieren.

Übersicht

Die optimale Trennebene



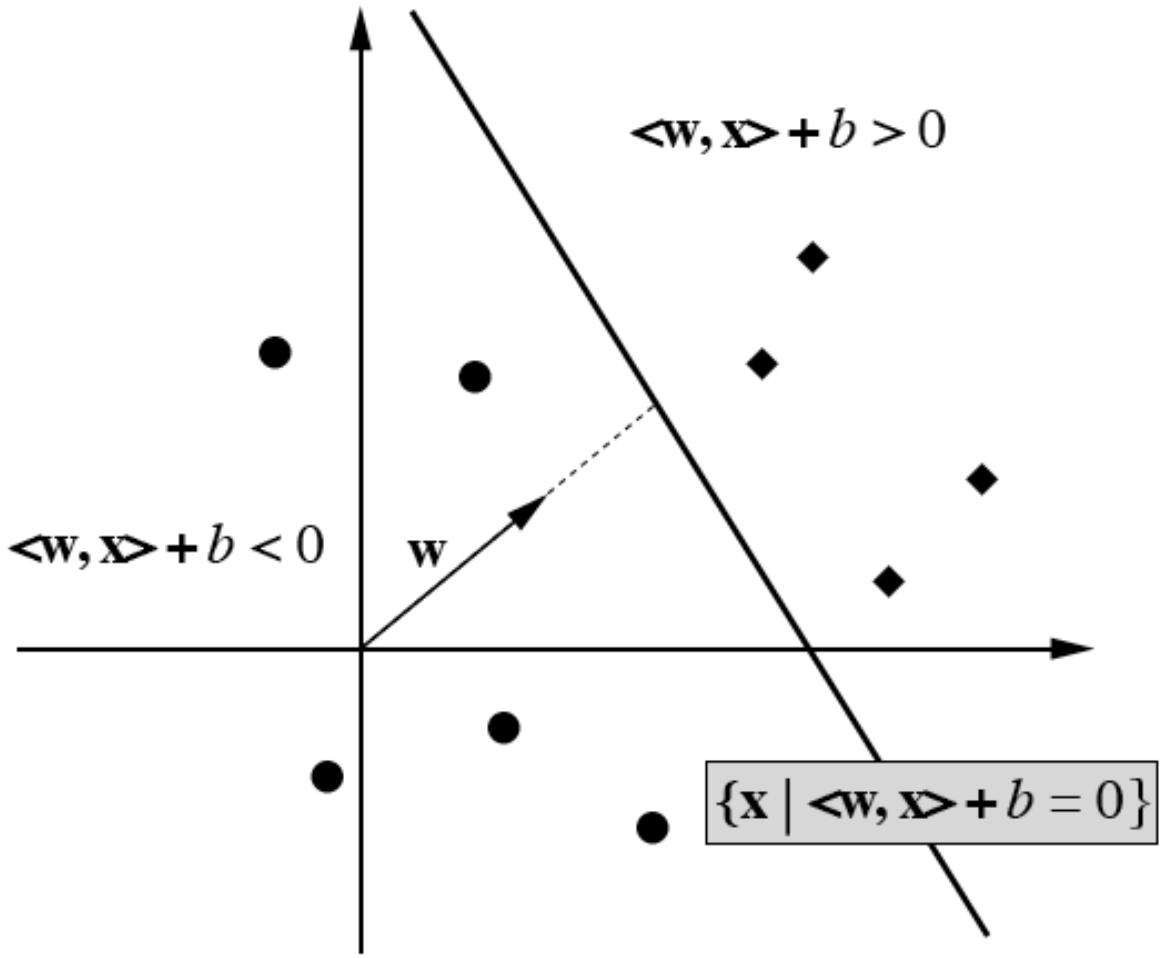




...which one?

Der Supportvektoralgorithmus sucht die Trennebene, die die Trennbreite für die gegebenen Trainingsdaten maximiert.

Wiederholung: Trennebenen



[B.Schölkopf]

Skalierungsinvarianz: für $c \neq 0$ beschreiben

$$\{\mathbf{x} | \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b = 0\}$$

und

$$\{\mathbf{x} | \langle c\mathbf{w} \cdot \mathbf{x}_i \rangle + cb = 0\}$$

dieselbe Ebene.

Kanonische Form (bzgl. Trainingsmenge S): wähle c so, daß

$$\min_{\mathbf{x}_i \in X} |\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b| = 1$$

gilt.

Trennbreite

Vorzeichenbehafteter Abstand eines Punktes zur Hyperebene (geometric margin):

$$\gamma_i = \frac{\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b}{\|\mathbf{w}\|}$$

Trennbreite einer Hyperebene $(\mathbf{w}, b)$ bzgl. einer Trainingsmenge S (margin):

$$\gamma = \min_{i=1..\ell} \frac{|\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b|}{\|\mathbf{w}\|},$$

d.h. der Abstand der zur Trennebene am nächsten liegenden Datenpunkte.

Bei einer kanonischen Trennebene gilt für die nächstliegenden Datenpunkte:

$$\gamma_i = \frac{1}{\|\mathbf{w}\|} \qquad \text{falls} \qquad y_i = 1$$

$$\gamma_i = -\frac{1}{\|\mathbf{w}\|} \qquad \text{falls} \qquad y_i = -1$$

Die nächstliegenden Datenpunkte heißen Supportvektoren (engl. to support: halten, unterstützen), da nur sie allein die Lage der optimalen Trennebene bestimmen.

Auffinden der optimalen Trennebene

- Bei einer optimalen Trennebene in kanonischer Form gilt für die Supportvektoren $\gamma_i = \pm\frac{1}{\|\mathbf{w}\|}$, d.h. es gibt einen Freiraum der Breite $\frac{2}{\|\mathbf{w}\|}$ um die Ebene herum, in der sich keine Datenpunkte befinden.
- Um die optimale Trennebene zu finden, müssen wir also $\frac{2}{\|\mathbf{w}\|}$ möglichst groß machen, d.h. $\|\mathbf{w}\| = \sqrt{w_1^2 + w_2^2 + \ldots + w_i^2 + \ldots} = \sqrt{\langle \mathbf{w} \cdot \mathbf{w} \rangle}$ minimieren.
- Da $\sqrt{\cdot}$ eine monoton steigende Funktion ist, kann man genauso gut $\langle \mathbf{w} \cdot \mathbf{w} \rangle$ minimieren, was mathematisch einfacher ist.
- Gleichzeitig muß sichergestellt sein, daß die Trainingsmenge korrekt getrennt wird, wir stellen die korrekte Trennung der Trainingsbeispiele als Nebenbedingung der Minimierung.

Supportvektoralgorithmus als Optimierungsproblem

SVM-Optimierung

Für einen gegebenen Trainingsdatensatz
$S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_\ell, y_\ell)\} \subseteq (X \times Y)^\ell$

- minimiere
$$f(\mathbf{w}) = \langle \mathbf{w} \cdot \mathbf{w} \rangle, \ \mathbf{w} \in \mathbb{R}^d$$
- unter ℓ Nebenbedingungen, $i = 1, \ldots, \ell$
$$\begin{aligned} \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b &\geq 1 \quad \text{für} \quad y_i = 1 \\ \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b &\leq -1 \quad \text{für} \quad y_i = -1 \end{aligned}$$

  oder kürzer
$$y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1$$

Es handelt sich also um ein quadratisches Optimierungsproblem mit linearen Ungleichheitsbedingungen.

Übersicht

Wiederholung: Optimierung unter Nebenbedingungen

Optimierungsproblem mit Nebenbedingungen

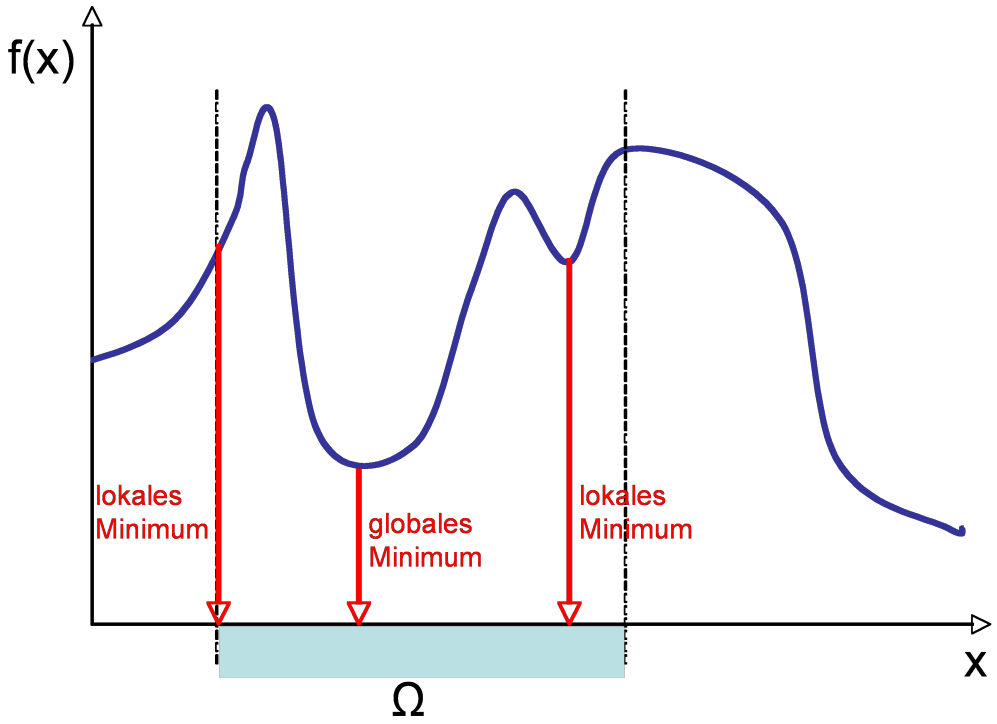
$$\text{minimiere} \quad f(\mathbf{x}), \quad \mathbf{x} \in \Omega$$

unter den Nebenbedingungen:

$$\begin{aligned} g_i(\mathbf{x}) &\leq 0, \quad i = 1 \ldots k \quad \text{(Ungleichheitsbedingungen)} \\ h_i(\mathbf{x}) &= 0, \quad i = 1 \ldots m \quad \text{(Gleichheitsbedingungen)} \end{aligned}$$

- Durch die Funktionen $g_i(\mathbf{x})$ und $h_i(\mathbf{x})$ wir die Wahl von $\mathbf{x}$ auf die Bereiche von Ω eingeschränkt, in denen die Gleichheits- und Ungleichheitsbedingungen erfüllt sind.
- Eine Ungleichheitsbedingung $g_i(\mathbf{x})$ heißt aktiv, wenn am Minimum $\mathbf{x}^*$ $g_i(\mathbf{x}^*) = 0$ gilt (d.h. gerade noch erfüllt ist), ansonsten heißt sie inaktiv.
- Ω heißt der zulässige Bereich des Optimierungsproblems, ein Punkt in diesem Bereich heißt zulässig (”feasible”).

Lokale und globale Optima



- Ein zulässiger Punkt $\mathbf{x}^* \in \Omega$ heißt globales Minimum, wenn für alle $\mathbf{x} \in \Omega$ gilt:

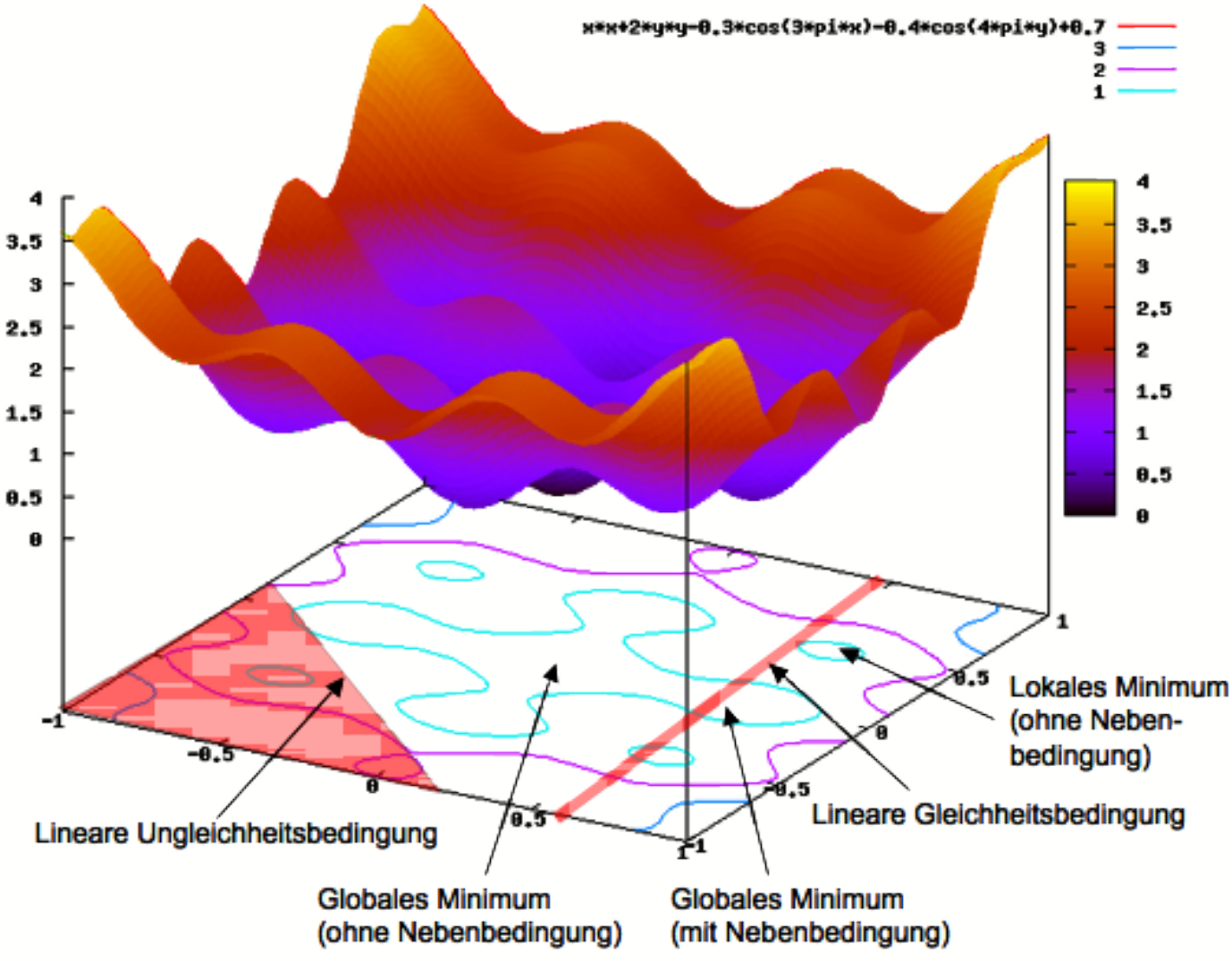
$$f(\mathbf{x}^*) \leq f(\mathbf{x})$$

- Ein zulässiger Punkt $\mathbf{x}^* \in \Omega$ heißt lokales Minimum, wenn es einen Radius $r > 0$ gibt so daß für alle $\mathbf{x} \in \Omega$ mit $\|\mathbf{x}^* - \mathbf{x}\| < r$ gilt:

$$f(\mathbf{x}^*) \leq f(\mathbf{x})$$

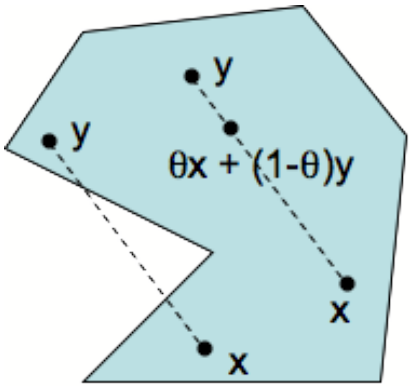
Die Lösung des Optimierungsproblems ist also ein globales Optimum.

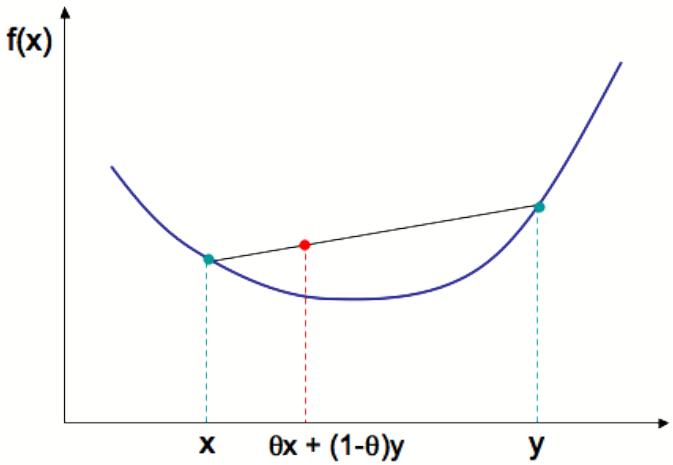
Beispiel: Optimierung unter Nebenbedingungen



Konvexe Mengen und konvexe Funktionen

Die Optimierungsaufgabe wird erheblich erleichtert, wenn es sich um ein konvexes Optimierungsproblem handelt. Wichtige Konzepte sind hierbei konvexe Mengen und Funktionen.





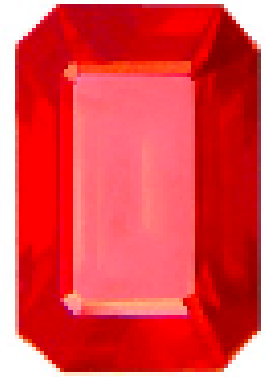
Eine Menge X heißt konvex, wenn für alle Punkte $\mathbf{x}, \mathbf{y} \in X$ und beliebiges $0 \leq \theta \leq 1$ gilt:

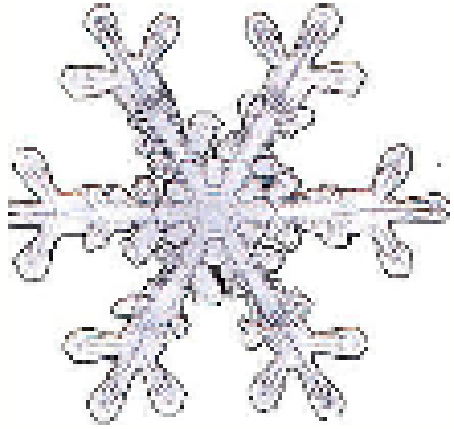
$$\theta\mathbf{x} + (1-\theta)\mathbf{y} \in X.$$

Intuitiv: Eine Menge ist konvex, wenn alle Punkte zwischen zwei Elementen der Menge ebenfalls Elemente der Menge sind.
Eine Funktion f heißt konvex, wenn für alle Punkte $\mathbf{x}, \mathbf{y}$ und beliebiges $0 \leq \theta \leq 1$ gilt:

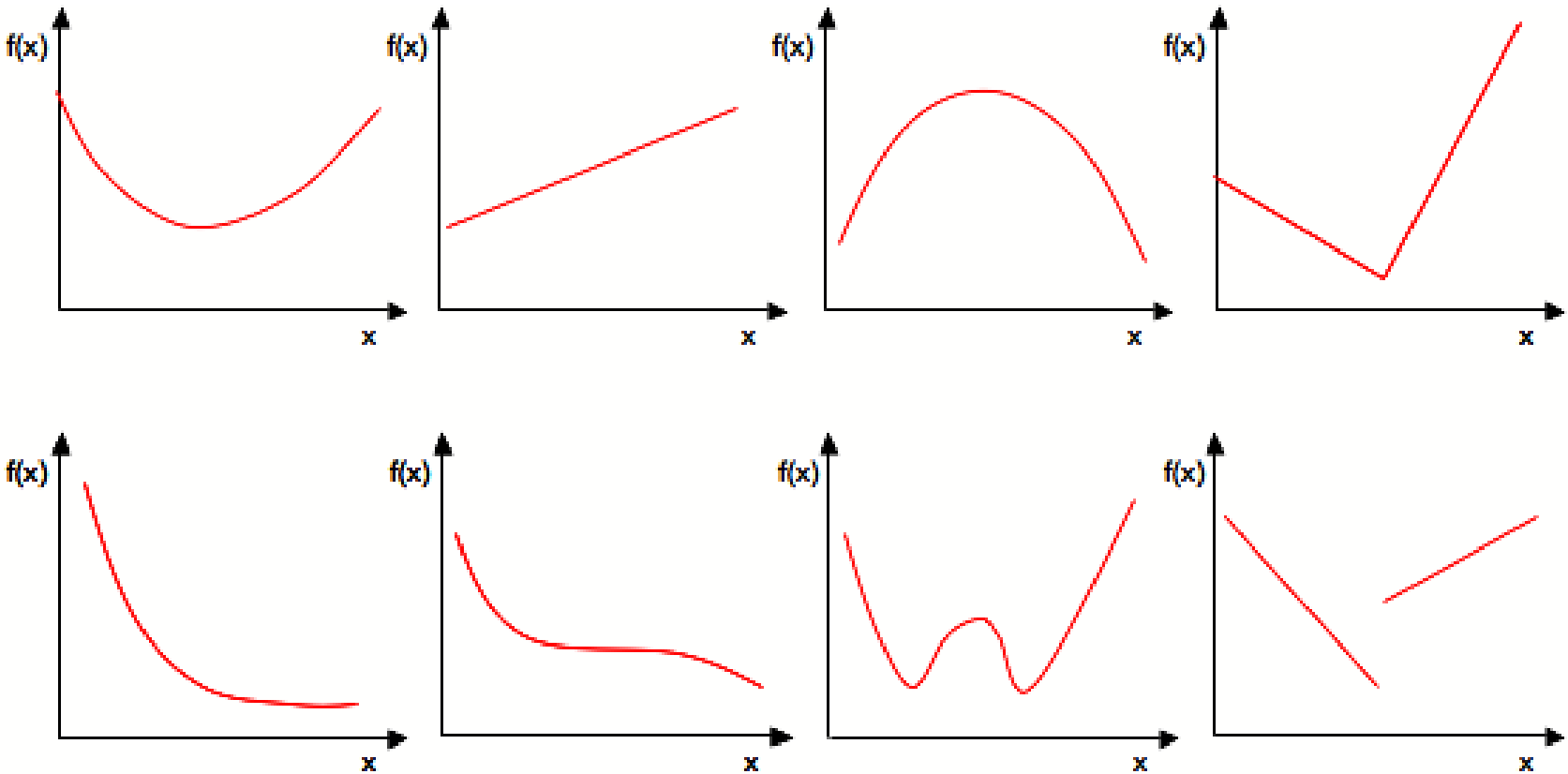
$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y})$$

Quiz: Welche Mengen sind konvex?





Quiz: Welche Funktionen sind konvex?



Eigenschaften von konvexen Mengen und Funktionen

- Schnittmengen von konvexen Mengen sind ebenfalls konvex.
- Vereinigungsmengen von konvexen Mengen müssen nicht konvex sein.
- Die Summe zweier konvexer Funktionen ist ebenfalls konvex.
- Das positive Mehrfache einer konvexen Funktion ist ebenfalls konvex.
- Differenzen zweier konvexer Funktionen müssen nicht konvex sein.
- Wenn f eine konvexe Funktion ist, dann ist die Menge $\{\mathbf{x}|f(\mathbf{x}) \leq 0\}$ konvex.

Beispiele für konvexe Funktionen

- Lineare Funktionen $\langle \mathbf{w} \cdot \mathbf{x} \rangle + b$ sind konvex.
  Beweis:

$$\begin{aligned} f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) - (\theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}) &= \\ \langle \mathbf{w} \cdot (\theta\mathbf{x} + (1-\theta)\mathbf{y}) \rangle + b - (\theta(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b) + (1-\theta)(\langle \mathbf{w} \cdot \mathbf{y} \rangle + b) &= \\ \langle \mathbf{w} \cdot (\theta\mathbf{x} + (1-\theta)\mathbf{y}) \rangle + b - (\langle \mathbf{w} \cdot (\theta\mathbf{x} + (1-\theta)\mathbf{y}) \rangle + b) &= 0 \end{aligned}$$

- Die Funktion $f : x \to x^2$ ist konvex.
  Beweis:

$$\begin{aligned} f(\theta x + (1-\theta)y) - (\theta f(x) + (1-\theta) f(y) &= \\ (\theta x + (1-\theta)y)^2 - (\theta x^2 + (1-\theta)y^2) &= \\ -\theta(1-\theta)(x^2 - 2xy + y^2) = -\theta(1-\theta)(x-y)^2 &\leq 0 \end{aligned}$$

- Die Funktion $f : \mathbf{x} \to \langle \mathbf{x} \cdot \mathbf{x} \rangle$ ist konvex.
  Beweis: folgt aus der Konvexität von x^2 und $\langle \mathbf{x} \cdot \mathbf{x} \rangle = x_1^2 + x_2^2 + \ldots$.

Konvexe Optimierungsprobleme

Satz

Wenn Ω eine konvexe Menge und f eine konvexe Funktion ist, dann ist jedes lokale Minimum von f ein globales Minimum über Ω.

- Bei konvexen Problem kann also die Lösung mit lokalen Methoden (z.B. Gradientenabstieg) gefunden werden, ohne daß man ganz Ω absuchen muß.
- Achtung: Dieser Satz sagt nicht, daß für jede konvexe Menge und jede konvexe Funktion ein solches Minimum existiert.
- Unter bestimmten Bedingungen kann die Existenz eines globalen Minimums garantiert werden (z.B., wenn Ω nichtleer und kompakt und f stetig ist).

Ist die SVM-Optimierung konvex?

SVM-Optimierung

- minimiere $f(\mathbf{w}) = \langle \mathbf{w} \cdot \mathbf{w} \rangle$, $\mathbf{w} \in \mathbb{R}^d$
- unter ℓ Nebenbedingungen, $i = 1, \ldots, \ell$

$$y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1$$

- Die Funktion $f : \mathbf{w} \to \langle \mathbf{w} \cdot \mathbf{w} \rangle$ ist konvex (s. vorher) und stetig.
- Die Nebenbedingungen $y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) - 1$ sind lineare Funktionen und damit ebenfalls konvex.
- Damit ist der durch jede Nebenbedingung definierte zulässige Bereich Ω eine konvexe Menge (s. Eigenschaften), in diesem Fall ein Halbraum im $\mathbb{R}^d$.
- Der zulässige Bereich ist eine Schnittmenge von konvexen Mengen und damit selbst konvex.

Merkmale der SVM

- Der SVM-Algorithmus ist ein konvexes Optimierungsproblem, d.h. jedes lokale Minimum ist eine Lösung.
- Die Zielfunktion ist stetig und der zulässige Bereich kompakt (falls es eine nichtleere Schnittmenge gibt, d.h. falls das Problem linear trennbar ist) $\Rightarrow$ die Existenz einer Lösung ist garantiert.
- Im Vergleich zu traditionelleren konnektionistischen Ansätzen (z.B. MLPs) stellt die Konvexität einen entscheidenden Vorteil da, da man bei der Optimierung nicht in lokalen Minima hängen bleiben kann und die Optimierung immer das globale Minimum findet (”Off-the-shelf”-Algorithmus).
- Durch das Kriterium der maximalen Trennbreite ist die Wahrscheinlichkeit für eine Überanpassung gering, unabhängig von der Dimensionalität der Eingangsdaten $\Rightarrow$ auch auf hochdimensionale und nichtlineare Probleme anwendbar.